Research Paper | Education | E-ISSN No : 2454-9916 | Volume : 9 | Issue : 10 | October 2023

IMAGE OR VIDEO DESCRIPTION GENERATOR

*Pikki Lovaraju[1], T. Kishore Kumar[2], V. Gopi[3], T. Rama Kotaiah[4], T. Lalas Maruthi[5], Y. Suresh[6]

[1,2,3,4,5] B.Tech , Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur, AP (Corresponding Order)[1]

[2]Assistant Professor, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur, AP

ABSTRACT

Image or Video Description Generator is challenging because it requires the model to understand the visual content of the image or video, as well as the ability to generate natural language descriptions. One common approach for this is to use a combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. CNNs are well-suited for extracting visual features from images and videos, while LSTMs are well-suited for modeling sequential data, such as text. The LSTM is trained on a dataset of images or videos with paired textual descriptions. During training, the LSTM learns to predict the next word in the description given the current word and the visual features of the image or video. Once the model is trained, it can be used to generate descriptions for new images or videos. To do this, the model is simply given the image or video as input, and it outputs a textual description.

KEYWORDS: Image, Video, CNN, LSTM, Neural Networks, Description

1. INTRODUCTION

Image or Video Description Generator uses the concepts of natural language processing and computer vision to predict the given image/video and describe it in the English like language. By using Natural Language processing(NLP) is also used detect the objects but it is a old process and also the results predict by using NLP is not accurate. To overcome this, we use CNN, LSTM to predict the better results.

While human beings are able to do it easily, it takes a strong algorithm and a lot of computational power for a computer system to do so. CNN analyses the visual imaginary by scanning them from left to right and top to bottom and extracting relevant features. Finally, it combines all the parts for image classification. This project will also elaborate on the functions and structure of the various Neural networks involved. Generating image or video descriptions is an important aspect of Computer Vision and Natural language processing.CV2 library is used to convert the video into series of frames i.e., set of images and This Model aims to detect different objects found in an image, recognize the relationships between those objects and generate description.

2.MATERIALS AND METHODS:

2.1. Dataset:

Image Datasets: Utilize widely recognized image datasets such as MS COCO (Common Objects in Context), ImageNet, and Flickr32k, containing a large number of images with corresponding 4 textual descriptions for each image.

Video Datasets: For video description generation, video datasets containing annotated frames or videos are used. Popular options include YouTube-8M and ActivityNet.

2.2. Data Preprocessing:

For images, resize and normalize the images, and perform data augmentation techniques.

For videos, extract frames and apply image preprocessing techniques to the frames.

2.3. Convolutional Neural Network (CNN):

One kind of deep learning algorithm that works especially well for tasks involving picture recognition and processing is the convolutional neural network (CNN). Convolutional, pooling, and fully connected layers are some of the layers that make it up.

Convolutional Layer:

In a convolutional neural network (CNN), a type of layer known as a convolutional layer applies a collection of filters to the input data in order to produce feature maps that highlight the presence of features that have been discovered in the input. After each convolution operation, a CNN applies Rectified Linear Unit(ReLU) activation function transformation to the feature map, introducing nonlinearity to the model. he ReLU activation function is differentiable at all points except at zero. For values greater than zero, we just consider the max of the function.

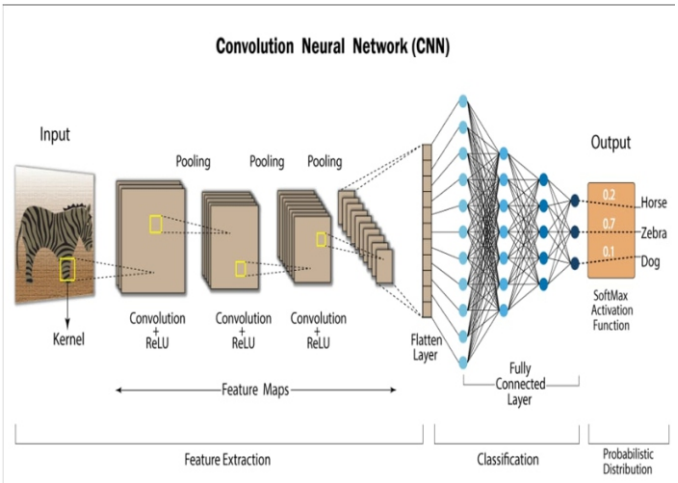

Figure 1: Overview of CNN

Pooling:
Pooling is just reducing the size of the image without losing the features that we found with convolution. For example, a MaxPooling method will take in a shape of a matrix and return the larger value in that range. By doing this we can compress the image without losing the important features of this image.

Flattening:
Flattening is nothing but converting a 3D or 2D matrix into a 1D input for the model this will be our last step to process the image.

Fully connected layer:
It takes the input from the previous layer and computes the final classification or regression task.

Output Layer:
The output from the fully connected layers is then fed into a logistic function for classification tasks like sigmoid or softmax which converts the output of each class into the probability score of each class.

2.4. Long Short Term Memory(LSTM):
LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) that can process sequential data and learn long-term dependencies. LSTM networks have a special structure that consists of a cell, an input gate, an output gate, and a forget gate. These gates regulate the flow of information into and out of the cell, allowing the network to remember or forget previous state

Forget Gate:
The information that is no longer useful in the cell state is removed with the forget gate. Two inputs i.e, input at particular time and previous cell output are fed to the gate and multiplied with weight matrices followed by the addition of bias.

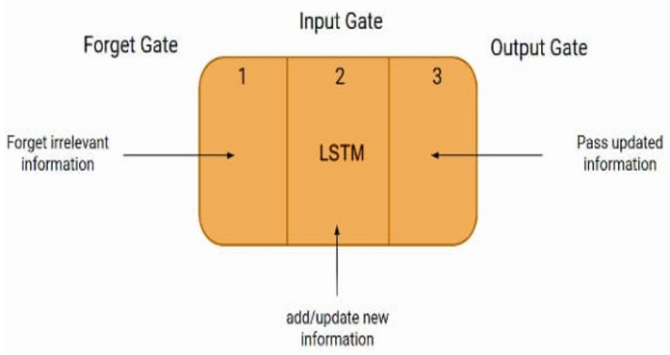

Figure 2: Overview of Long Short-Term Memory

Input Gate:
The addition of useful information to the cell state is done by the input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered similar to the forget gate using previous inputs and. Then, a vector is created using the tanh function that gives an output from -1 to +1, which contains all the possible values and At last, the values of the vector and the regulated values are multiplied to obtain useful information.

Output Gate:
The task of extracting useful information from the current cell state to be presented as output is done by the output gate. First, a vector is generated by applying the tanh function on the cell. Then, the information is regulated using the sigmoid function and filtered by the values to be remembered using inputs. At last, the values of the vector and the regulated values are multiplied to be sent as an output and input to the next cell.

2.5. Complete Architecture:
In this project, we have used VGG16 model which is a Standard convolutional Neural network has 16 layers used for image recognition and reduces the size of the image i.e., dimensions and extract the feature called feature map. The preprocessing of data is done by feeding input data into the VGG16 model application of Keras running on top of TensorFlow.

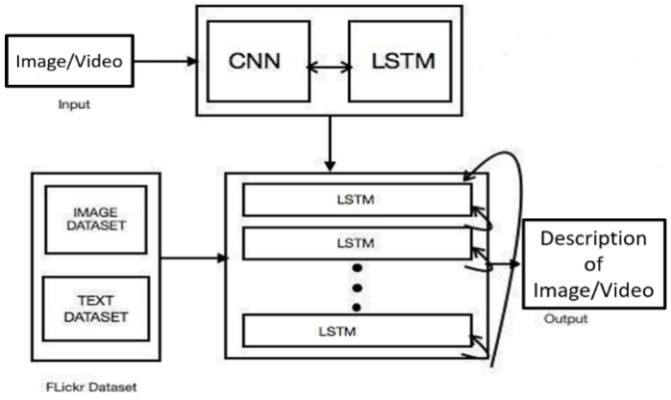

Figure 3: Overview of System Architecture

3. RESULTS AND DISCUSSIONS:
The results may be classified based on input as follows:
1. Image based Results
2. Video based Results

3.1. For Image:

Figure 4: Result of Sample Input-1

The process involves taking an image as input, performing preprocessing to standardize dimensions and normalize pixel values, and extracting features using a pre-trained Convolutional Neural Network (CNN). These features are then processed by a Long Short-Term Memory (LSTM) network to generate textual descriptions. The model is trained and fine-tuned using datasets of image-text pairs to minimize disparities. Once trained, it can accept images and produce human-readable descriptions, enhancing accessibility, content understanding,

and discoverability for various applications, including accessibility tools, content indexing, and recommendation systems.

3.2. For Video:

```
Generated Description for the given video:
['a large brown bear walking across a river']
```

Figure 5: Result of Sample Input-2

When the user uploaded the input as video, then it checks whether it is video format or not. If not, it will show a message that - 'Please upload valid video format'. If the uploaded format is correct, then using cv2 library the video is converted into collection of frames and those frames will be input as images which will generate the textual description.

4. CONCLUSION:

In this paper, we have learned and designed a technique of Image or Video Description Generator which will respond to User with description based on an image or video. The Image Based Model extracts features of an image and the Language based model translates the features and objects extracted by image based model to a natural sentence. Image based model uses CNN whereas Language Based model used LSTM.

The workflow is Data gathering followed by Pre-processing, Training model and Prediction. The ultimate purpose of an Image/Video description generator is to improve the social media platforms as well as in image indexing and for visually impaired persons with automated generated description.

REFERENCES

1. CS771 Project Image Captioning by Ankit Gupta, Kartik Hira, Bajaj Dilip.
2. Every Picture Tells a Story: Generating Sentences from Images. Computer Vision ECCV (2016) by Farhadi, Ali, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hocken-maier, and David Forsyth
3. Automatic Caption Generation for News Images by Yansong Feng, and Mirella Lapata, IEEE (2013).
4. Image Caption Generator Based on Deep Neural Networks by Jianhui Chen, Wenqiang Dong and Minchen Li, ACM (2014).
5. Show and Tell: Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE (2015).
6. Image2Text: A Multimodal Caption Generator by Chang Liu, Changhu Wang, Fuchun Sun, Yong Rui, ACM (2016).
7. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions by Sepp Hochreiter.
8. Where to put the Image in an Image Caption Generator by Marc Tanti, Albert Gatt, Kenneth P. Camilleri.
9. Sequence to sequence -video to text by Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Institute for Articial Intelligence, Kaiserslautern, Germany.

5.1 WEB REFERENCES:

1. https://towardsdatascience.com/a-guide-to-image-captioninge9fd5517f350
2. https://ieeexplore.ieee.org/document/8276124
3. https://machinelearningmastery.com/develop-a-deep-learningcaption-generation-model-in-python/
4. https://medium.com/@raman.shinde15/image-captioning-with-flickr8k-dataset-bleu4bcba0b52926
5. https://blog.clairvoyantsoft.com/image-caption-generator-535b8e9a66ac
6. https://www.matecconferences.org/articles/matecconf/abs/2018/91/matecconf_eitce2018_01052/matecconf_eitce2018_01052.html
7. https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/